Submitted to Bioinformatics
draft material -- please do not cite

BIOINFORMATICS

Vol. 00 no. 0 2005, pages 1–5
doi:10.1093/bioinformatics/bti283

Databases and ontologies

Predictive integration of Gene Ontology-driven similarity and functional interactions

Francisco Azuaje[1,*], Haiying Wang[1], Olivier Bodenreider[2] and Alban Chesneau[3]

[1]School of Computing and Mathematics, University of Ulster, UK., [2]National Library of Medicine, National Institutes of Health., USA, [3] Structural Genomics Group, EMBL-Grenoble, France.

Received on ; revised on…..; accepted on ….

Advance Access publication . . .

ABSTRACT

Motivation: The inference of functional networks of genes relies on the integration of multiple data and knowledge bases. Moreover, there is a need to develop methods to automatically incorporate prior knowledge to support the prediction and validation of novel functional associations. One such important knowledge source is represented by *The Gene Ontology* (GO)™ and the many model organism databases of gene products annotated to the GO. We investigated quantitative relationships between the GO-driven similarity of genes and their functional interactions by analyzing different types of associations in *S. cerivisiae* and *C. elegans*.

Results: This study demonstrates that interacting genes (including regulatory and protein-protein interactions) exhibit significantly higher levels of GO-driven similarity in comparison to random pairs of genes defined as negative interactions. Significant associations were identified using annotations from the three GO hierarchies, but it was confirmed that the Biological Process hierarchy may provide more reliable results for all of the types of interactions and organisms studied. Statistical analyses indicated that GO-driven similarity represent a relevant and relatively accurate resource to support prediction of functional networks in combination with other resources.

Availability: Supplementary information, including data sets generated are available at:
http://ijsr32.infj.ulster.ac.uk/~e10110731/GO-Inter

Contact: fj.azuaje@ieee.org

1 INTRODUCTION

The reliable prediction of functional networks of genes may be achieved by integrating multiple types of data sources, such as gene expression, phylogenetic profiles and high-throughput protein-protein interaction experiments. This is necessary because such individual sources may be considered as *weak prediction models* due to their limitations in terms of predictive accuracy and coverage. Several studies have reported significant links between different types of genomic data sets, as well as techniques, e.g. machine learning, to combine them and improve prediction quality for relatively simple model organisms, such as yeast (Jansen *et al.*, 2003; Lee *et al.*, 2004). Furthermore, it is crucial to integrate prior knowledge resources, such as annotation databases and literature, for not only building advanced functional classifiers, but also to assist in the validation of technique-independent predictions, e.g. detecting potential spurious associations.

The Gene Ontology (GO)™ is one such source of prior knowledge, which is becoming the *de facto* standard for annotating gene products (The Gene Ontology Consortium, 2001). It has been proposed as a gold standard to assess the quality of several classification systems using, for example, expression data (Al-Shahrour *et al.*, 2004). Moreover, information extracted from GO-driven annotation databases have been applied for making *de novo* predictions of gene function in relatively simple organisms (King *et al.*, 2003).

The GO has been proposed as a tool for measuring similarity between genes. Previous research showed significant relationships between semantic similarity of pairs of genes and their sequence-based similarity (Lord *et al.*, 2003). Also we have evaluated relevant quantitative relationships between GO-driven similarity and gene expression correlation (Wang *et al.*, 2004). GO-driven clustering algorithms based on such approaches have been recently reported (Wang *et al.*, 2005). Moreover, they have provided the basis for developing tools that may facilitate the identification of relevant partitions from clustering, using, for example, GO-driven cluster validity indices (Bolshakova *et al.*, 2005).

Prior to the integration of a predictive resource, *Res*, is first necessary to assess its predictive relevance and reliability in relation to data sets of known positive and negative interactions (Jansen and Gerstein, 2004). In this case the hypothesis to prove is: Can information extracted from *Res* be applied to distinguish pairs of interacting genes (positives) from those that have not shown evidence to be interacting (negative)?

The application of GO-driven annotation information to support the prediction of functional networks of genes has not been rigorously investigated. Jansen *et al.* (2003) integrated different genomic data sets including annotations derived only from the GO Biological Process hierarchy to predict protein-protein (PP) interactions. The *GO-driven similarity* of a pair of genes was used as an indicator of PP interactions in yeast. Between-gene similarity was calculated by identifying the set of GO terms shared by the two sets of annotations. For a given database of protein pairs, the total number of protein pairs sharing the same set of annotations was used as an estimator of similarity. Thus, the lower this frequency value, the more similar the gene pair under consideration. They found that lower term counts were correlated with a higher likelihood of finding two proteins in the same complex. Nevertheless, such a similarity assessment approach does not fully exploit relevant topological and information content features that may be

*To whom correspondence should be addressed.

useful for meaningfully estimating between-gene similarity. In some cases genes that are annotated to closely related but distinct GO terms may actually exhibit no similarity.

Using annotations from the three GO hierarchies: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC), we sought to assess relationships between the GO-driven similarity of a pair of genes and their functional interactions. This study aimed to investigate the feasibility of applying GO-driven similarity to support the prediction of functional interactions of genes, including physical and regulatory interactions, in *S. cerevisiae* and *C. elegans*. Two key questions addressed were: a) Can GO-driven similarity be applied to estimate the functional coupling of genes, such as gene expression co-regulations and other physical and non-physical interactions and b) Can such a knowledge be used in combination with other resources to improve the prediction process? Our hypothesis is that the GO-driven similarity of a pair of genes may be used as a relevant indicator of functional interaction.

The following section describes the data sets analyzed: 1) A data set of annotated co-regulatory interactions from yeast, 2) An extensive, high-quality functional gene network for yeast, and a 3) high-quality PP interaction data set from *C. elegans*. This is followed by a description of the methods applied to measure similarity using GO annotations, its links to the identification of interacting pairs of genes and a statistical assessment of the predictions. Results for the three data sets are presented. The final section discusses the main contributions and limitations of this study, potential applications and future research.

2 MATERIALS

2.1 Data sets

Gene co-regulation in *S. cerevisiae* (CoReg)
This data set originated from a comprehensive collection of annotated regulons compiled by Simonis *et al.* (2004). Their data set comprised more than 1400 pairs of gene-factor associations retrieved from the TRANSFAC (Wingender *et al.*, 2000) and aMAZE (van Helden *et al.*, 2001) databases and literature searches. More than 13000 pairs of co-regulated genes were then extracted from these data. These pairs comprised the CoReg reference data set analyzed in this investigation.

Functional network of yeast genes (FunNet)
This data set was obtained from an extensive, high-quality functional gene network investigated by Lee *et al.* (2004). Unlike CoReg data set, FunNet comprises different types of functional associations: Mediated and non-mediated by physical interaction, i.e. PP, regulatory, etc. This network was inferred by integrating diverse, high-quality functional data sets: mRNA coexpression, gene-fusions, phylogenetic profiles, literature co-citation and protein interaction experiments, with the Kyoto Encyclopedia of Genes and Genomes (KEGG) database used as the gold standard. A sub-sample of 19216 pairs of genes representing the most reliable interaction predictions were analyzed in this study (supplementary paper offers additional information).

PP interactions in *C. elegans* (PPInt):
This data set represents another level of complexity, in which 860 PP interactions, including a few self-interactions, were obtained from the *Worm Interactome* (WI5) map. The selected data set, from now on referred to as PPInt, contains the highest-confidence, published interactions from WI5 (Li *et al.*, 2004).

2.2 The GO

The GO hierarchies: MF, BP and CC encode annotation terms that describe the role played by a gene product, the biological goals to which a gene product contributes and the cellular localization of the gene product respectively. Such vocabularies of annotation terms (one for each hierarchy) and their relationships are represented by *directed acyclic graphs*, in which each annotation term may represent a "child node" of one or more "parent nodes". There are two types of child-to-parent relationships in the GO: "is a" and "part of" types. The first type is defined when a child annotation term is a subclass of a parent term. The second type is used when a parent has the child as its part. This study takes advantage of both types of links as justified elsewhere (Lord *et al*., 2003). The GO comprises annotation terms supported by different types of evidence codes, such as the TAS (Traceable Author Statement) and IEA (Inferred from Electronic Annotation) codes. The TAS code refers to annotations supported by peer-reviewed papers. In contrast, IEA annotations are based on predictions automatically obtained from sequence similarity searches, which have not been reviewed by curators. Detailed information on GO-driven annotation databases, their development and evidence codes supported is available at www.geneontology.org. The reader is also referred to (Azuaje *et al.*, 2005) and (Wang *et al.*, 2004) for an introduction to some of the predictive data analysis applications of the GO.

2.3 GO annotation databases

The pairs of interacting genes in each of the above data sets were described by their GO annotations. *IEA* annotations were excluded from these analyses due to their lack of reliability. The March 2005 database releases of the *Saccharomyces Genome Database* (SGD) and *WormBase* (WB) provided the GO annotations for these data sets, which are available at www.godatabase.org. All of the CoReg interacting pairs (13412) were composed by genes in which both genes had at least one GO annotation from all hierarchies. FunNet had 19003 pairs of interacting genes with both genes assigned to at least one GO annotation under all GO hierarchies. In the PPInt data the numbers of interacting pairs of genes in which both genes were described by at least one GO annotation were 188, 296 and 77 pairs under the MF, BP and CC hierarchies respectively. The Supplementary Section offers a more detailed description of the data sets analyzed.

3 METHODS

3.1 GO-driven similarity

In order to estimate the similarity of a pair of genes, g_k and g_p, one must first understand how to calculate the similarity between the terms belonging to the sets, A_k and A_p, used to annotate these

genes. Different methods, known as *information-theoretic approaches*, have been previously studied to measure ontology-driven similarity (Lord *et al.*, 2003; Azuaje *et al.*, 2005). Unlike traditional *edge-counting techniques*, these methods are based on the assumption that the more information two terms share in common, the more similar they are. The *Lin's similarity model*, for example, has shown to produce both biologically meaningful and consistent similarity predictions (Lord *et al.*, 2003; Wang *et al.*, 2004) in comparison to related approaches. Given terms, $c_i \in A_k$ and $c_j \in A_p$, the between-term Lin's similarity is defined as:

$$sim(c_i, c_j) = \frac{2 \times \max_{c \in S(c_i, c_j)} [\log(p(c))]}{\log(p(c_i)) + \log(p(c_j))} \quad (1)$$

where $S(c_i,c_j)$ represents the set of parent terms shared by both c_i and c_j, 'max' represents the maximum operator, and $p(c)$ is the probability of finding a child of c in the annotation database being analyzed. It generates normalized similarity values between 0 and 1. Thus, given a pair of gene products, g_k and g_p, with A_k and A_p comprising m and n terms respectively, the between-gene similarity, $SIM(g_k, g_p)$, may be defined as the average inter-set similarity between terms from A_i and A_j:

$$SIM(g_k, g_p) = \frac{1}{m \times n} \times \sum_{c_i \in A_k, c_j \in A_p} sim(c_i, c_j) \quad (2)$$

where $sim(c_i,c_j)$ may be calculated using (1). This and other similarity approaches, as well as their relationships with sequence-based similarity and co-expression, have been investigated in (Lord *et al.*, 2003) and (Wang *et al.*, 2004). The results reported in this paper are based on the calculation of gene similarity based on (1) and (2). The last section discusses some of the limitations of this technique.

3.2 Linking GO-driven similarity and functional interactions

GO-driven similarity values were calculated for all the annotated pairs in the data sets described in Section 2. These data represented our sets of true positive interactions, which were statistically analyzed to show significant relationships with GO-driven similarity. In order to illustrate such links, similarity values from these sets of true positive interactions were compared to similarity values generated by a set of negative interactions, i.e. pairs of genes not showing evidence of interaction. Thus, a set of "non-interacting genes" was produced as follows. For a given data set, **P**, comprising M true positive interactions, a set **N**, with M negative interactions was built by randomly pairing genes from **P**. Moreover, the resulting sets were verified to ensure that newly formed pairs were not included in **P**. This process is also equivalent to the idea of randomly permutating the lists of GO annotations, A_k, $1 \leq k \leq M$, describing the genes in **P** to form a new set **N**. One has to take into account that some of the pairs included in **N** may actually be false negatives (true positives) and this might influence the comparisons performed. However, at least with regard to the data sets analyzed (evidence available) this could not be demonstrated. The final section of this paper further discusses this factor. The resulting data sets **N** represent a valid approximation of counter-examples, which are essential to explore potential associations between functional interactions and GO-driven similarity. Furthermore, the random effects and variability linked to this data sampling procedure may be reduced by generating K independent **N** sets. These K sets may be then analyzed as an aggregated set, **N'**, consisting of K x M pairs of (non-interacting) genes.

Fundamental relationships between GO-driven similarity and the existence/absence of functional interactions were estimated by comparing similarity values exhibited by **P** versus values observed in **N'**. Their similarity value distributions for each of the problems described in Section 2 and for all of the GO hierarchies were analyzed. Differences between **P** and **N'** were summarized by estimating their respective mean similarity values. The significance of their differences was tested by applying Student's t-Test. The relevant null hypothesis tested was that these mean similarity values originated from the same sample, i.e. there are no significant differences between these mean values. This relatively simple task provided key insights into relationships between the degree of similarity of pairs of genes and the likelihood that these genes are functionally interacting.

After identifying significant differences, the capacity of GO-driven similarity to predict functional interactions (as a single predictive source) was analyzed. Given a similarity value, $SIM(g_k, g_p)$, and a pre-defined *predictive similarity threshold* value, *GOS-Th*, genes g_k and g_p are said to be an interacting pair (positive interaction) if $SIM(g_k,g_p) \geq GOS\text{-}Th$. Some of these predictions will obviously be false positive interactions. Therefore, the next task was to estimate the rate of false predictions. This was done by estimating the proportion of the number of interactions that would occur by chance to the number of pairs correctly predicted as positive interacting pairs. This represents the ratio of the number of false positive predictions, *Ro*, to the number of true positives predictions, *R*. *Ro* was estimated using the mean number of interacting pairs obtained from the K data sets, **N**, i.e. the total number of interactions observed in **N'** divided by K. This is related to the problem of estimating the *decisive false discovery rate*, which has shown to be a robust and conservative estimator of the probability, P, of detecting spurious associations (Bickel, 2005). Thus, when this rate of false predictions, P, is closer to '1', the stronger the evidence to suggest few or no true positive interactions. That is, lower P values indicate stronger evidence to support the validity of the positive interactions detected by the GO-driven similarity method. P values were calculated for the data sets described above using different *GOS-Th* values. This analysis allows one to have a better idea about how many false positive predictions may potentially be made when applying the GO-driven similarity method as a single prediction model.

4 RESULTS

The analysis tasks described above were implemented with $K = 10$, 10 and 100 for the CoReg, FunNet and PPInt data sets respectively. This relatively small number of randomly generated sets,

N, was selected due to computing power limitations. Similar results were obtained for other comparisons involving other *K* sets, **N**. Moreover, the relatively large number of gene pairs included at least in the first two data sets should contribute to the reduction of the bias and variability of the estimations. The supplementary section provides additional information and results.

4.1 Results from CoReg

Table 1 summarizes the differences between the sets **P** and **N'** with regard to their mean similarity values. The high *t* values obtained suggest significant differences ($p < 0.001$) for all GO hierarchies. Histograms are used in Fig. 1 to illustrate differences of the similarity value distributions for **P** and one of the data sets **N** in connection to the BP hierarchy. Interacting pairs of genes in general exhibit higher similarity values than non-interacting pairs. Similar trends were obtained for other **N** sets and hierarchies (see Supplementary Section). This indicates the feasibility of applying GO-driven similarity to support the distinction of co-regulated from non-co-regulated pairs of genes. Fig. 2 shows the estimated probabilities, *P*, that such predictions are false as a function of the predictive threshold, *GOS-Th*.

Table 1. CoReg data set. Differences between interacting and random, non-interacting pairs of genes in terms of their GO-driven similarity. SE: standard error of the estimated mean.

GO Hierarchy	True Positives (Mean±SE)	Random Pairs (Mean±SE)	*t* values
MF	1.8E-01 ± 2.4E-03	1.1E-01 ± 5.4E-04	2.8E+01
BP	2.3E-01 ± 2.3E-03	1.3E-01 ± 4.5E-04	4.3E+01
CC	3.1E-01 ± 2.3E-03	2.4E-01 ± 6.2E-04	2.9E+01

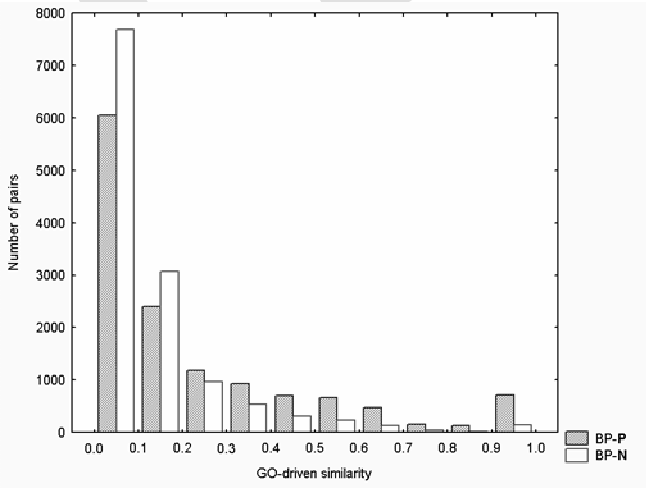

Fig. 1. CoReg: Distribution of similarity values from **P** and one of the **N** data sets under the BP hierarchy.

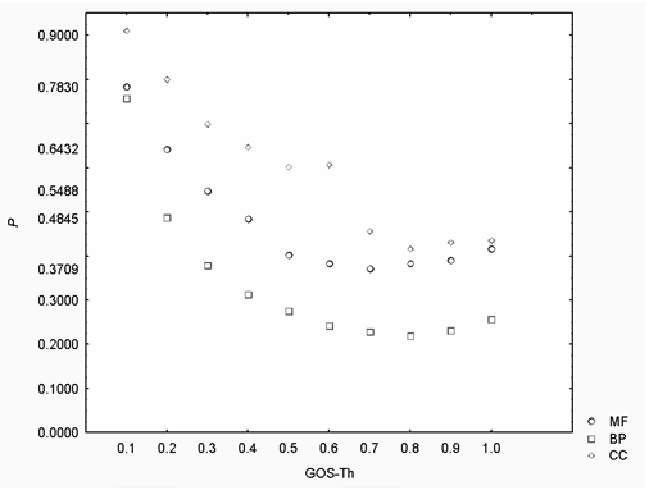

Fig. 2. CoReg: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

4.2 Results from FunNet

Table 2 summarizes the differences between the sets **P** and **N'** in terms of their mean similarity values. The high *t* values obtained suggest significant differences ($p < 0.001$) for all GO hierarchies. Fig. 3 illustrate differences between the similarity value distributions from **P** and one of the data sets **N** with regard to the BP hierarchy. Interacting pairs tend to exhibit higher similarity values than non-interacting pairs. Similar properties were obtained for other **N** sets and hierarchies (see Supplementary Section). This suggests the feasibility of using GO-driven similarity to help to distinguish interacting from non-interacting pairs of genes (including physical and non-physical interactions). Fig. 4 shows the estimated probabilities, *P*, that such predictions are false as a function of the predictive threshold, *GOS-Th*.

Table 2. FunNet: Differences between interacting and random, non-interacting pairs of genes in terms of their GO-driven similarity. SE: standard error of the estimated mean.

GO hierarchy	Interacting pairs (Mean±SE)	Random pairs (Mean±SE)	*t* values
MF	4.9E-01 ± 3.3E-03	2.0E-01 ± 8.4E-04	8.8E+01
BP	5.4E-01 ± 2.7E-03	2.8E-01 ± 6.9E-04	9.2E+01
CC	5.7E-01 ± 2.4E-03	3.4E-01 ± 6.4E-04	9.0E+01

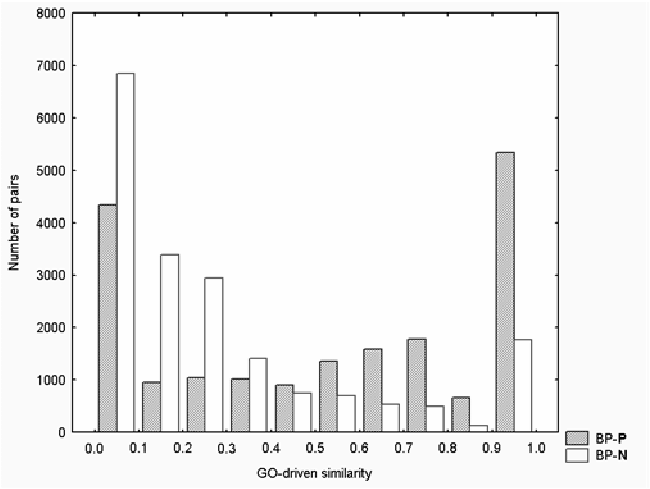

Fig. 3. FunNet: Distribution of similarity values from **P** and one of the **N** data sets under the BP hierarchy.

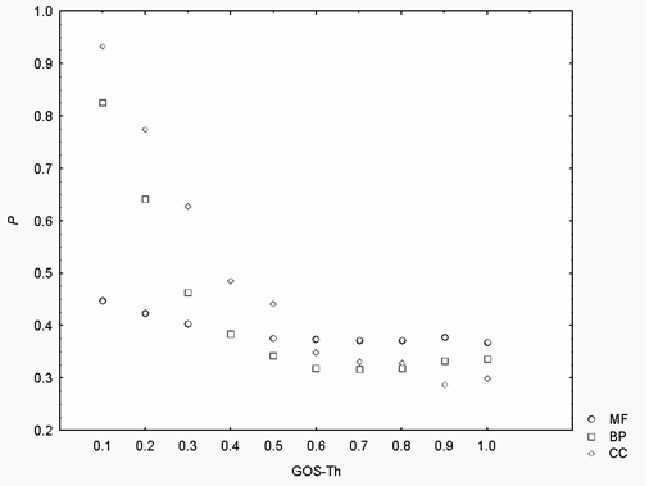

Fig. 4. FunNet: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

4.3 Results from PPInt

Table 3 summarizes the differences between the sets **P** and **N'** with regard to their mean similarity values. The high *t* values obtained also suggest significant differences ($p < 0.001$) in connection to all GO hierarchies. Fig. 5 depicts differences of the similarity value distributions from **P** and one of the data sets **N** regarding the BP hierarchy. Again the interacting pairs tend to produce stronger similarity values than non-interacting pairs. Similar trends were obtained for other **N** sets and hierarchies (see Supplementary Section). This may suggest the potential of GO-driven similarity to assist in the differentiation of interacting and non-interacting pairs of proteins in more complex organisms. Fig. 6 presents the estimated probabilities, *P*, that such predictions are false as a function of the predictive threshold, *GOS-Th*.

Table 3. PP-Int: Differences between interacting and random, non-interacting pairs of genes in terms of their GO-driven similarity. SE: standard error of the estimated mean.

GO hierarchy	Interacting pairs (Mean±SE)	Random pairs (Mean±SE)	*t* values
MF	2.7E-01 ± 2.3E-02	1.6E-01 ± 1.7E-03	5.0E+00
BP	1.9E-01 ± 1.4E-02	1.2E-01 ± 8.1E-04	4.9E+00
CC	3.9E-01 ± 4.6E-02	2.0E-01 ± 3.6E-04	4.0E+00

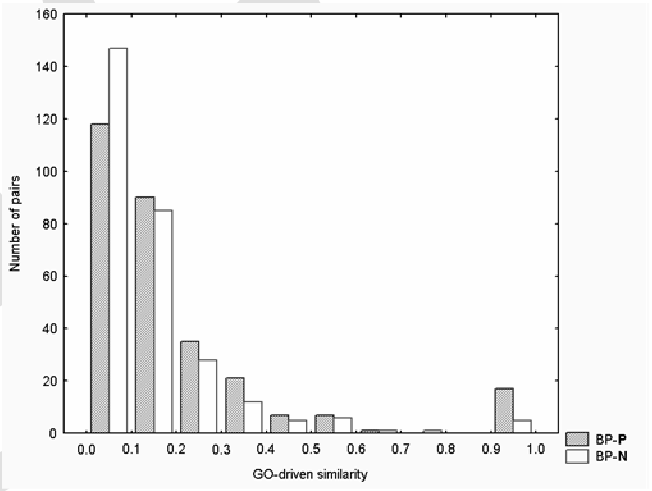

Fig. 5. PP-Int: Distribution of similarity values from **P** and one of the **N** data sets under the BP hierarchy.

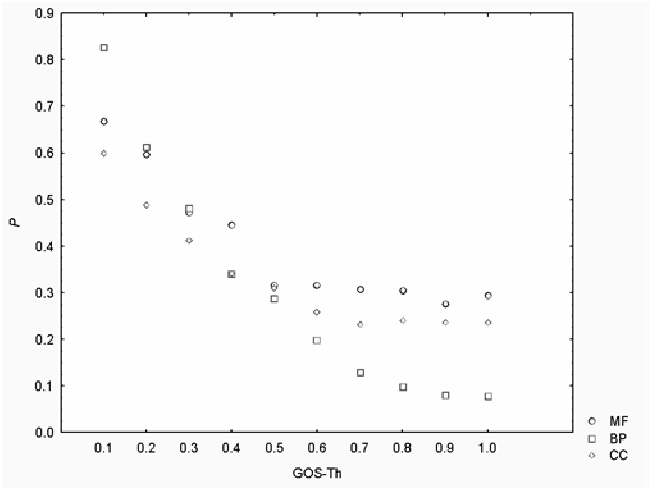

Fig. 6. PP-Int: Rate of false positive predictions, *P*, as a function of the *GOS-Th* for all GO hierarchies. *P* estimates the probability of predicting spurious associations.

5 DISCUSSION AND CONCLUSIONS

Relationships between GO-driven similarity, based on an information theoretic approach, and different levels of functional interaction were investigated. Three complexity levels were explored:

Co-regulation in *S. cerevisiae*, a more comprehensive set of functional associations (including both physical and non-physical interactions) in the same organism and a smaller set of PP interactions in *C. elegans*.

We focused our investigation on previously published, high-quality annotated interactions, which represented the reference data sets for this study. The GO terms used to describe each pair of genes did not include electronically inferred annotations. We concentrated on a GO-driven similarity assessment approach that has previously shown to be strongly related to sequence-based similarity and gene co-expression (Lord *et al.*, 2003; Wang *et al.*, 2004). This paper demonstrated significant relationships between the GO-driven similarity shown by a pair of genes and their interaction. This pattern was remarkably observed under all hierarchies. This supports the hypothesis that the GO-driven similarity of pair of genes may be applied to support the prediction of functional interactions (including co-regulatory and PP interactions). Additional experiments, which are summarized in the Supplementary Section, also suggested that the degree of GO-driven similarity may be consistent with interaction likelihood scores of pairs of genes as reported by Lee *et al.* (2004) based on a comprehensive, integrative prediction strategy. Our research does not of course suggest that this approach is sufficient and necessary to detect relevant interactions. Similarly, we did not aim to argue that it may represent a more effective prediction model than existing approaches. However, this investigation offered evidence to motivate the application of this functional similarity measure as a complementary predictive resource of functional interaction. This, in combination with other sources, such as gene co-expression and different interaction prediction models, may support more accurate and biologically-meaningful predictions. Integrative prediction models such as those reported by (Lee *et al*., 2004) and (Jansen *et al.*, 2003) may be benefited from incorporating this knowledge-based source.

Figs. 1, 3 and 5 highlight two important protein groups. The first represent protein pairs that are not similar based on their GO terms. These proteins may be true negatives, but also they may represent pairs that are actually interacting even though their available annotations are unrelated. This could be the case for proteins without GO annotations or for those involved in processes not properly described by this ontology. The second set corresponds to similar proteins pairs in relation to their GO-driven annotations. As demonstrated above, such similarity may offer strong evidence for the presence of functional interaction. This may also be merged with other post-genomic sources of evidence, such as genome-wide in-situ hybridization, to improve the detection of false positive interactions. Thus, the GO-driven similarity approach may also complement experimental approaches to determining false positives. In the case of metazoan organisms, e.g. *C. elegans*, the GO-driven similarity approach is much more difficult to assess as the function of a protein can be related to its tissue- or organ-specific expression patterns. It would be important to integrate gene expression and tissue localization information to complement GO-driven similarity. Moreover, it might be possible to infer tissue localization from GO annotations. This dimension, which is not considered in unicellular organisms, underlies the complexity of this prediction task and the importance of implementing integrative, module-based approaches to interactome prediction.

P.H Lee and D. Lee (2005) recently integrated ontology-driven similarity information as part of their *modularized network learning method* (MONET). They first recognized modules of interrelated genes using gene expression correlation and MIPS (Munich Information center for Protein Sequences database) annotations. *Bayesian networks* were then inferred from the detected modules that successfully predicted relevant gene regulation networks in yeast. Ontology-driven similarity was required to aid in the identification of clusters of genes on the basis of their MIPS annotations. Between-gene similarity was estimated using the between-term Resnik's method (1995), which is also an information-theoretic approach. But unlike Lin's method, Resnik's method generates unnormalized similarity values ranging from 0 to infinity. Moreover, previous research has shown that Lin's technique may outperform Renisk's and other information-theoretic approaches (Lin, 1998). For example, Lin's method may generate similarity values highly correlated with human assessments of similarity in different application domains. Between-gene similarity based on GO-driven Lin's method may reflect significant relationships with gene co-expression. Such relationships may be represented in a more consistent and meaningful fashion in comparison to Resnik's approach (Wang *et at.*, 2004). Wang *et al.* (2005) proposed a GO-driven hierarchical clustering method based on Lin's technique, which recognized significant functional modules relevant to several responses to stimuli in yeast. Their method may complement P.H. Lee and D. Lee's method (2005) for the detection of functional modules based on GO-annotations. A recent study on global prediction of regulatory networks in yeast found that more 12% of genetic interactions included genes with identical GO annotations (Tong *et al.*, 2004). It also found that over 27% of the interactions comprised similar annotations sets based on a conservative estimate of similarity, which approximated the degree of annotation overlaps. Our investigation provides further evidence of the potential of GO-driven similarity information to facilitate the prediction of functional interactions. Moreover, we have shown that these relationships may go beyond the regulatory level and may support applications involving uni- and multi-cellular organisms.

The similarity value distributions, significant differences and the potential to reduce the probability of detecting spurious associations encourage further investigations. Moreover, we believe that, even when significant results were obtained, our assessment may actually be under-estimated. This is because many of the pairs of genes included in the randomly-generated sets ("true negatives") might indeed be part of more comprehensive collections of true positive interactions not included in this study. Some of them might also become true positive interactions in the future with the emergence of new experimental and validated evidence. This factor also suggests that the differences and relationships identified could be stronger. Nevertheless, the relatively large amount of interacting pairs included in the Co-Reg and FunNet data sets may contribute to the reduction of such noise sources and possible bias.

One of the major challenges for predicting interaction maps is to remove false-positives interactions. False positives predictions can be caused by technical or biological factors. The former have no biological meaning and come from technical limitations such as the number and the strength of phenotypic tests used for two-hybrid screenings or purification steps realized for complex identification by mass spectrometry. The latter ones may originate from proteins that actually interact but which are not expressed in the same tissues or organs. For example, it was estimated that about 25% to 50% of the genome-wide protein-protein interaction pre-

dictions reported in many high-profile publications actually represent false positive interactions (Edwards *et al.*, 2002).

This investigation also estimated probability values, *P*, that offered relevant insights into these relationships. These indicators may help us to further assess the predictive ability of the GO-driven similarity method to detect valid interactions for different prediction similarity thresholds, *GOS-Th*. Figs. 2, 4 and 6 suggest that in general the larger the *GOS-Th*, the lower the probability of making false positive predictions. But it also highlights the fact that many of the false positive interaction predictions show high level of similarity (including the maximum similarity). This may also be explained by the difficulties in creating exact true negative data sets. As expected, the greater the *GOS-Th* value, the lower the number of positive interactions (both true and false) made. But in some cases the reduction in the number of false positive interactions was visibly smaller than the reduction of the number of true positive interactions (Fig. 2, for example). This means that the results obtained cannot be used as conclusive evidence to indicate that higher *GOS-Th* should necessarily produce more accurate predictions. They suggest that there is a tendency to reduce the number of false positive interactions by applying more rigorous thresholds. The results obtained also confirm that the higher the *GOS-Th* the more limited the predictive coverage of the model, i.e. the higher the possibility of missing true positive interactions. This property is perhaps more visible in the CoReg and FunNet data sets. In the case of the PP-Int data set, a stronger inverse proportional relationship between *P* and *GOS-Th* was observed under the BP and CC hierarchies. Predictors based on annotations from the MF hierarchy showed to be the most unreliable for this data set. The lowest *P* value was obtained when the predicted interactions were based on the maximum *GOS-Th*.

The lack of clearer, more regular response patterns may also be explained by the difficulties in building more reliable data sets representing true negative interactions as discussed above. Figs. 2, 4 and 6 confirm that, in principle, it would be possible to implement more accurate predictive models with higher *GOS-Th* values, but at the risk of detecting a considerable number of false positive interactions and of reducing predictive coverage. Predictions based on annotations from the BP hierarchy are indicated as the most reliable and accurate predictions for the three data sets analyzed. These and the results shown in Tables 1 to 3 cannot be considered as surprising findings. However, they highlight the potential of using GO-driven similarity as an alternative *weak prediction model*, which may complement other weak predictive resources, e.g. gene co-expression and high-throughput interaction identification techniques. It also opens opportunities to incorporate prior knowledge to support the automated assessment or validation of predictions derived from large-scale studies.

Another aspect that deserves further research is the design of between-term similarity assessment methods that can reflect biological features in a more intuitive and meaningful way. For instance, one would expect that $SIM(g_i,g_j) = 1$, when $g_i = g_j$ or when A_i represents the same set of annotations A_j. However, this is not always the case when the method described by (2) is applied. It will define, for instance, $SIM(g_i,g_j) = 0.5$, for $g_i = g_j$ when A_i is described by $m > 1$ annotations. This factor may also have contributed to an under-estimation of the significance of the results. Alternative methods should be evaluated to address these inconsistencies, which are particularly critical when dealing with data sets involving self-interactions. In order to address such a limitation we are currently evaluating an alternative between-gene similarity approach that selectively aggregates maximum between-term similarity values (Azuaje *et al.*, 2005).

As part of future research we will analyze other functional databases related to the organisms considered in this paper as well as others such as mouse. We will incorporate other knowledge resources, such as KEGG databases, to further assess the applications and implications of GO-driven similarity assessment for supporting large-scale interactome prediction studies.

REFERENCES

Al-Shahrour,F., Diaz-Uriarte,R., and Dopazo,J. (2004) Fatigo: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578-580.

Azuaje, F., Wang, H., and Bodenreider, O. (2005) Ontology-driven similarity approaches to supporting gene functional assessment. In *Proc. Of The Eighth Annual Bio-Ontologies Meeting*, Michigan, 25 June, http://bio-ontologies.man.ac.uk/

Bickel, D.R. (2005) Probabilities of spurious connections in gene networks: application to expression time series. *Bioinformatics*, **21**, 1121-1128.

Bolshakova,N., Azuaje, F., and Cunningham,P. (2005) A knowledge-driven approach to cluster validity assessment. *Bioinformatics*, **21**, 2546-7.

Edwards, A.M., Kus, B., Jansen R., *et al.* (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes.*Trends in Genet*ics, **10**, 529-36

Jansen, R. and Gerstein, M. (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol*, **7**: 535-45.

Jansen, R., Yu, H., Greenbaum, D., *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449-53.

King, O. D., Foulger, R. E., Dwight, S. S., *et al.* (2003) Predicting gene function from patterns of annotation. *Genome Research*, **13**, 896--904.

Lee, I, Date S.V., Adai, A. and Marcotte, E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555-8.

Lee, P.H., and Lee, D. (2005) Modularized learning of genetic interaction networks from biological annotations and mRNA expression data. *Bioinformatics*, **21**, 2739-47

Li, S., Armstrong, C.M., Bertin, N. and *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540-3.

Lin, D. (1998) An information-theoretic definition of similarity. in *Proc. of 15th International Conference on Machine Learning*, San Francisco, 296-304.

Lord, P., Stevens, R., Brass, A. and Goble, C. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275--1283.

Prokisch, H., Scharfe, C., Camp II, D.G., *et al.* (2004) Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biology*, **2**, 796-803.

Resnik, P. (1995) Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th International Joint Conference on Artificial Intelligence*, Montreal, 448-453.

Simonis, N., Wodak, S.J., Cohen, G.N. and van Helden, J. (2004) Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics*, **15**, 2370-9.

Tong, A.H. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808-813.

The Gene Ontology Consortium (2001) Creating the gene ontology resource: Design and implementation. *Genome Research*, **11**, 1425-1433.

van Helden,J., Rios,A.F. and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyad. *Nucleic Acid Res.*, **28**, 1808-18.

Wang,H., Azuaje,F., Bodenreider,O., and Dopazo,J. (2004) Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proc. of IEEE 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, La Jolla, CA, USA, 25-31.

Wang,H., Azuaje,F., and Bodenreider,O. (2005) An ontology-driven clustering method for supporting gene expression analysis. In *Proc. of the 18th IEEE International Symposium on Computer-Based Medical Systems,* in press.

Wingender,E., Chen,X., Hehl,R., *et al.* (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acid Res.*, **28**, 316-319.